1 Introduction

First: Overview and Arrangements
What this course is about?

1.1 Structured Documents
Review of basic notions

Goals of the Course

- To familiarize with the most important models and languages for
  - manipulating
  - representing
  - transforming and
  - querying
  structured documents (or XML)
- "Core XML processing technology"
  - little about specific XML applications or commercial systems

NOT an Exhaustive Survey

- Emphasis on active formalisms (for describing processes on documents) instead of describing documents/data
- Bias in selecting course topics:
  - estimated usefulness/value
    - centrality (implying longer-lasting value)
    - maturity: Stable specifications? Existing implementations?
  - Lecturer up-to-date?

Motivation?

- Academic interest on models of information processing
- Practical relevance: "eBusiness" is HOT!

Preliminary Outline

1 Introduction
  Overview and Arrangements
  1.1 Structured Documents
2 Document Instances and Grammars
  2.1 Trees and their Grammars
  2.2 Review of XML basics: DTDs, Namespaces, Schemas
3 Programmatic Manipulation of Structured Documents (XML APIs)
  3.1 SAX
  3.2 DOM
### Preliminary Outline (2)

4 Styling Structured Documents I
   4.1 Essentials of Cascading Style Sheets
5 Transforming Structured Documents
   5.1 Addressing: XPath
5.2 XSLT
6 Styling Structured Documents II: XSL
7 XML Web-Site Architectures
8 XML wrapping (or translating data to XML)
9 Querying Structured Documents
   9.1 Region Algebra and agrep
9.2 XML Query Languages

### Methodological Goals

- Some central professional skills
  - consulting of technical specifications
  - experimenting with SW implementations
- Ability to think…?
  - to find out relationships
  - to apply knowledge in new situations
- (*"Pidgin English" for scientific communication)

### Administration

- An elective graduate-level (laudatur) special course
  - suitable for all specialisation lines (esp. CS/SWE)
- 3 cu (≈120 hours of work)
- Lectures Mar 11 - May 8, Microteknia MT2
  - Lecturer: Pekka.Kilpelainen@uku.fi
- Assistant: Sami.Komulainen@uku.fi

### Administration: Exercises

- Exercises Mar 20 - May 15(?), MT2
  - essential for familiarising with the subject
  - mainly normal homework assignments, solutions discussed in class
  - 1 or 2 groups, depending on attendance
- + a few (1-3) "mini-projects"
  - reading and summarising tasks?
  - hands-on experimentation?
  - to be handed in to lecturer
  - credited like other exercises (scaled based on quality by a factor in [0, 1.5])

### Administration: Grading

- Course examination on Wed, May 22, in the Great Lecture hall (SL)
  - minimum of 50% of exam points to pass the course
- Grade = \((32\times\text{Exam}/\text{MaxExam} + 12\times\text{HomeWork}/\text{MaxHomeWork} - 8)/3\)
- Opportunity to retake the exam
  - June 12 (≥ 50% to pass, grade with/without homework credits)

### Material

- No single textbook
- Reports, articles
- Course home page
  - lecture notes, exercises, reference material, announcements, …
- Recommended (but not required) text:
Background Check

- Basic knowledge of structured documents and document standards
  - Course "Document standards"?
  - HTML?
- Programming languages and concepts
  - OO programming, Java?
  - Unix/Linux / Windows?
- Formal language theory
  - Theory of Computation / "Ohjelmoinnin ja laskennan teoria"?
  - regular expressions, automata?
  - context-free grammars, parse trees?

1.1. Structured Documents

- **Document:**
  - a structured representation of information on some medium (= message)
  - normally for a human reader
  - memos, manuals, articles, books, …
  - also application-to-application messages
  - EDI (electronic data interchange)
  - "prose-oriented XML" vs "data-oriented XML"
  - possibly non-permanent, dynamically generated
  - processable or conceivable as a unit
  - (a web page vs a web site)

Text-Based Documents

- We concentrate on textual or text-based documents
  - character data major constituent of information content
  - as opposed to, say multimedia documents
- Next: Presentation vs Structure

Presentation vs Structure

- Presentation informs the human reader about the meaning of text and the role of its parts
- **Markup:** indicating the presentation or the meaning of different parts of text
  - originally hand-written annotations for the typesetter
  - nowadays primarily codes embedded in digital documents

Markup

- **Procedural markup**
  - formatting commands (start boldface, produce an empty line, indent 5 mm, …)
  - proprietary word processor formats, nroff, TeX, …
- **Descriptive or generic markup**
  - indicating the logical structure of text using chosen names
  - LaTeX: \begin{abstract} … \end{abstract}
  - HTML: <TITLE> … </TITLE>
- **Markup language**
  - a fixed set of markup notations (e.g. nroff, TeX, HTML, SVG, …)
Structured documents?

Most liberally, any document is structured (punctuation, words, sentences, fields, …) but especially descriptively marked-up documents … especially if they adhere to a rigorous specification of structure.

Structure in documents

- **Hierarchy** or nesting is ubiquitous – chapters of books, warnings in maintenance manuals, …
- **Linear order** essential in prose documents – less important in documents representing data objects
- **Hypertext** and cross-references
- We’ll be mainly dealing with manipulation of hierarchical, or tree-like document structures

Next: How these are modelled?